

RESEARCH ARTICLE

Predicting corporate financial distress based on quantitative and qualitative information of annual report using text mining techniques

Sohrab Osta^a, Mojtaba Alifamian^b, Mohammad Marfou^{c1}, Behrang Parsafard^d

^a Associate Professor of Accounting, University of Ilam, Ilam, Iran.

^b PhD in Accounting, Allameh Tabataba'ei University, Tehran, Iran.

^c Corresponding Author, Assistant Professor, Department of Accounting, Allameh Tabataba'ei University, Tehran, Iran.

^d PhD in Accounting, Allameh Tabataba'ei University, Tehran, Iran.

ARTICLE INFO

Keywords:

Financial Distress

Text Mining

Machine Learning

Python, TF-IDF

ABSTRACT

As much as a firm's life cycle may come with numerous opportunities to grow and prosper, some conditions may burden the firms with periods of decline, crisis, and loss. The modern era of global economy entails that more firms are established while more are also financially distressed than in the past. Hence, business failure and bankruptcy have increased due to globalization and the integration of national economies. Financial distress has thus posed serious challenges for the financial life of countries as its personal and social costs highlight its significance for many managers, policy-makers, stakeholders, and auditors. Therefore, the present study seeks to analyze unstructured data from the annual report and combine it with quantitative financial information to predict financial distress through text mining techniques and machine learning algorithms. For this purpose, using the Fuzzy Delphi Method, quantitative variables affecting financial distress were identified based on the opinion of experts. Their quantitative information was extracted from financial statements and then, the report of 100 companies listed in the period 2013 - 2023 were collected and after converting the corresponding PDF files to the Word files, using python programming language, data mining (including preprocessing, feature extraction, feature selection, etc.) was developed. The prediction results indicate the high predictive power of SVM model with radial kernel compared to other models. So that its ability to predict financial distress based on three quantitative, qualitative and integrated methods is 85, 91 and 92% respectively. The results of this research showed that instead of just paying attention to the numbers and the ratios derived from these numbers, the text mining technique can also be used for analysis and prediction, and by combining it with the results obtained from quantitative information, the financial distress of companies can be determined. Although the accuracy of predicting results from unstructured data compared to structured data with a predictability about 90% is less, knowing and using the capacity of this type of information is very important so many academics and practitioners can believe that quantitative disclosures alone are not effective for economic decision-making.

1. Introduction

New approaches to firms as socio-economic units imply that besides their financial functions, firms serve significant purposes in non-financial contexts, such as contributing to the job market and employment capacity, preserving or damaging the environment, affecting future generations, and transferring capital or liabilities to them, and other vital roles in their community. Hence, firms' financial distress and bankruptcy or health and profitability significantly impacts socio-economic indices on both micro and macro levels considering their financial and non-financial consequences. Therefore, predicting such effects could help raise the necessary awareness for appropriate reactions to the probable future conditions of the firms (especially from an economic viewpoint), as using such information would allow better employment of investment opportunities and resource allocation. Such measures would provide the necessary warnings regarding the occurrence of financial distress and thus allow firms to take appropriate measures and, secondly, enable creditors and investors to invest their resources in profitable opportunities by differentiating favorable chances to invest from unfavorable ones (Mehrani et al., 2005).

The bulk of studies on insolvency and bankruptcy of companies used reports, structured data, and quantitative disclosures in recent years. However, large amounts of information are now published through qualitative disclosures (e.g., textual reports and documents such as mandatory files and disclosures like K-10, Q-10, and K-8 forms, profit notices and other declarations, telephone conferences and financial

¹ Corresponding Author

E-mail address: marfoua@gmail.com

Received 19 May 2026; Received in revised form 20 July 2026; Available online 10 August 2026



media, news, analytical notes, financial reports of analysts, declarations and interpretations of legislators such as the regulations of the Stock Exchange Commission, macro economy news and emotions, online messages, and social media), which are all different forms of unstructured financial data (Guo et al., 2016). Despite the great importance of quantitative information from the financial reports of companies, this kind of information cannot solely provide users with a comprehensive image of the financial status of a company. Therefore, investors need to first decode and then process the quantitative information of financial reports. This is why the collectors, users, and legislators of accounting information have long been interested in gathering and presenting written information as well as the quantitative information of financial reports to better perceive this kind of information and the information that cannot be presented numerically. However, textual information now accounts for a larger portion of quantitative information in the financial reports of companies (Hájek, P., & Michalak, 2013).

Text mining is a well-known tool for decoding and analyzing the information content of reports on companies. Referring to the process of extracting interesting but non-evident knowledge from unstructured texts, text mining is also known as the smart analysis of text, textual data mining, and knowledge extraction from texts (Gupta & Lehal, 2009). However, this research method (text mining) has received little emphasis in Iran due to its complexity and the unavailability of suitable hardware resources. In other words, we have not found any studies on the prediction of insolvency through the text analysis of accounting reports in Iran.

On the other hand, various models have been proposed to predict distress so far, each with their capacities and limitations. Despite their relative simplicity, traditional models suffer from limitations such as normality, linearity, and predictive variables' independence. Novel methods such as support vector machines and artificial neural networks have thus been proposed over time to enhance prediction performance and address previous shortcomings, offering better generalizability and performance in terms of overall model accuracy.

The present work seeks to initially convert unstructured data from annual reports into structured data and develop a model to predict financial distress by integrating financial and textual data through machine learning algorithms, and then proceed to compare the accuracy of financial distress prediction based on textual, financial, and integrated textual and financial data. Another highlight of the present work is its selection of financial data for prediction model development. Variable selection is generally based on research literature, which has undermined consensus among scholars in achieving a general result regarding the predictive financial variables. Researchers' findings indicate that some proposed variables are irrelevant to firm bankruptcy and distress. Hence, contrary to other works in this field that have examined accounting data only, the present work incorporates financial, market, and other data types by inquiring expert opinions when modeling financial distress. The second highlight of this research is concerned with the fact that the two distinct concepts of bankruptcy and financial distress have been used in the past literature interchangeably, to the point that Article 141 of the Commercial Law has been compiled to identify bankrupt and distress firms (this article identifies firms that lose half of their capital due to losses as distress). The present work thus reviews the research literature on financial distress and eligible firms and proceeds to enter the modeling process. Based on the identified research gaps regarding the limitations of traditional financial indicators and the potential contribution of textual information, this study seeks to answer the following research questions:

RQ1: Can textual information extracted from corporate annual reports provide significant predictive power for identifying financially distressed firms?

RQ2: Does the integration of textual and financial information improve the performance of financial distress prediction models compared with models based solely on financial or textual information?

RQ3: Which machine learning algorithm provides the highest predictive performance for financial distress prediction when using financial information, textual information, and their integration?

2. Literature Review

2.1. Traditional Approaches to Financial Distress Prediction

The financial health and stability of firms have long been considered critical factors in economic decision-making and investment activities. The financial condition of a company reflects its ability to withstand internal and external risks and maintain operational continuity ((Ključnikov et al., 2022; Kovacova & Klietlik, 2018; Szymura, 2022). Consequently, accurate prediction of financial distress has attracted considerable attention from researchers, investors, creditors, and policymakers, as early identification of distress conditions can facilitate timely corrective actions and improve resource allocation.

Early studies on financial distress prediction primarily relied on financial statements and accounting ratios. Fitzpatrick (1932) was among the first researchers to employ financial indicators for predicting corporate insolvency. Subsequently, Merwin (1942) demonstrated that financially distressed and healthy firms differed significantly in terms of financial ratios. A major milestone in this field was introduced by Beaver (1966), who examined the predictive ability of financial ratios in identifying distressed firms. Later, Altman (1968) developed the well-known Z-score model, which became one of the most influential frameworks in bankruptcy prediction research.

Following these developments, statistical models such as logit and probit models were introduced to improve prediction accuracy. Ohlson (1980) and Zmijewski (1984) proposed influential probability-based approaches for bankruptcy prediction. Although these traditional models provided valuable insights and remain widely used, they rely heavily on historical financial information and are subject to limitations related to assumptions of linearity, variable independence, and the ability of accounting data to fully represent firms' future conditions.

2.2. Machine Learning Approaches in Financial Distress Prediction

With advances in computational capabilities and artificial intelligence, machine learning techniques have increasingly been applied to overcome the limitations of traditional statistical models. These approaches can capture complex and nonlinear relationships among predictive variables and provide improved generalization capabilities in financial distress prediction (Altman et al., 2017).

Previous studies have demonstrated the effectiveness of machine learning algorithms in bankruptcy and financial distress prediction. Mai et al. (2019) investigated the application of deep learning techniques and found that deep learning models improved insolvency prediction performance compared with conventional approaches. Sreedharan et al. (2020) compared several machine learning algorithms and reported that neural network-based models achieved superior predictive accuracy among the examined methods. Li et al. (2021) also demonstrated the potential of deep learning approaches for developing financial distress prediction models.

Recent studies have further emphasized the importance of using multi-source information and advanced learning algorithms. Karimpour et al. (2025) developed a deep learning-based model for financial distress prediction and reported high predictive accuracy. Huang et al. (2024) proposed a Variable Deep Financial Distress Prediction (VDFDP) approach using incomplete multidimensional data and showed that their proposed method significantly outperformed benchmark models. Similarly, Wang et al. (2025) demonstrated that incorporating financial and non-financial big data through deep learning frameworks can substantially improve financial distress prediction performance.

Despite these advances, many machine learning-based studies have continued to rely primarily on structured financial indicators. Therefore, although machine learning techniques improve prediction accuracy, the potential contribution of unstructured corporate information remains an important area requiring further investigation.

2.3. Textual Information and Text Mining in Financial Distress Prediction

Financial reports contain not only numerical information but also substantial qualitative disclosures that reflect managerial perspectives, expectations, risks, and future business conditions. The increasing availability of unstructured financial data, including annual reports, management discussions, financial news, and corporate disclosures, has created new opportunities for improving financial distress prediction models (Guo et al., 2016).

Text mining techniques provide a systematic approach for extracting meaningful patterns from unstructured textual information. Text mining refers to the process of discovering valuable and previously unknown knowledge from textual documents (Gupta & Lehal, 2009). However, analyzing financial texts is challenging because of the complexity of language and the need to identify informative patterns beyond simple word frequencies.

Prior studies have highlighted the predictive value of textual information in financial distress prediction. Li et al. (2010) identified different approaches for analyzing corporate language, including dictionary-based methods and statistical approaches. Loughran and McDonald (2010) argued that general sentiment dictionaries may not be fully suitable for financial texts due to the specific meaning of words in business contexts. Therefore, data-driven text analysis approaches have received increasing attention.

Among corporate disclosures, annual reports are particularly valuable because they provide comprehensive information about firms' past performance, current conditions, and future expectations. Barzegari Khaneghah et al. (2019) emphasized that annual reports can serve as an important source for evaluating and predicting firms' financial conditions.

Empirical evidence has confirmed the incremental value of textual information. Jo and Shin (2016) found that emotional information extracted from economic news affected bankruptcy prediction outcomes, with negative expressions associated with higher bankruptcy risk. Tang et al. (2020) showed that textual characteristics of managerial reports, including positive and negative word frequencies and text length, improved insolvency prediction accuracy. Hajiha and Saberi (2021) also found that frequently used words in audit reports contributed significantly to corporate financial distress prediction.

2.4. Integrated Financial and Textual Models in Financial Distress Prediction

Recent research has shifted from relying exclusively on either financial indicators or textual information toward integrated approaches that combine multiple information sources. These models aim to capture both quantitative financial conditions and qualitative signals embedded in corporate disclosures.

Liu and Jia (2025) proposed a BERT-Text CNN framework that integrates textual information embedding with deep feature extraction techniques for financial distress prediction. Using data from 4,784 listed companies during 2007–2022, their findings indicated that the proposed model achieved superior predictive performance compared with traditional financial indicators and other data fusion approaches.

Similarly, Nguyen et al. (2026) demonstrated that incorporating textual information from corporate disclosures alongside traditional financial indicators significantly improves bankruptcy prediction performance. Their findings showed that topic-based and sentiment-based textual features provide valuable early warning signals of financial distress and enhance the interpretability of financial risk prediction models.

Although previous studies confirm the usefulness of textual information and advanced machine learning techniques, several limitations remain. First, many studies examine either financial indicators or textual features separately, providing limited evidence regarding their comparative predictive contribution within a unified framework. Second, limited research has investigated the incremental value of integrating textual and financial information, particularly in emerging markets. Third, differences in financial distress definitions and variable selection procedures have created inconsistencies among previous studies.

Therefore, the present study addresses these gaps by developing and comparing three categories of financial distress prediction models: models based on financial information, models based on textual information extracted from annual reports, and integrated models combining both financial and textual features. Furthermore, machine learning algorithms are employed to identify the most effective approach for predicting financial distress.

3. Research Methodology

3.1 Research Design and Framework

The present study aims to develop an integrated model for predicting financial distress by combining financial and textual information using machine learning techniques. The research framework consists of five sequential stages: (1) extraction of quantitative variables, (2) achievement of expert consensus through the Fuzzy Delphi method, (3) data collection and identification of financially distressed firms, (4) text mining and feature extraction, and (5) model training and performance evaluation.

3.2 Data Collection and Sample Selection

The study population consists of companies listed on the Tehran Stock Exchange (TSE). To identify the determinants of financial distress, the quantitative variables affecting financial distress were first identified using the Fuzzy Delphi method based on expert opinions, and the corresponding financial information was extracted from the firms' financial statements.

Subsequently, the annual reports of 100 listed companies over the period 2013–2023 were collected from the Comprehensive Database of All Listed Companies (CODAL), the official database of the Securities and Exchange Organization of Iran. The final dataset consisted of approximately 1,100 firm-year observations. After converting the PDF reports into Microsoft Word documents, text mining procedures were implemented using the Python programming language, including text preprocessing, feature extraction, and feature selection. Companies with incomplete financial information or poor-quality reports that could not be properly converted into machine-readable text were excluded from the final sample.

3.3 Extraction of Quantitative Variables and Expert Consensus

An exploratory review of the relevant literature was first conducted to identify quantitative variables associated with financial distress in the Iranian economic and institutional environment. Subsequently, a research questionnaire was developed and evaluated for reliability and validity before being distributed among the expert panel.

The Fuzzy Delphi method was employed to achieve expert consensus regarding the factors influencing financial distress prediction. The experts were selected using purposive judgmental sampling based on their academic knowledge and professional experience in financial distress prediction and corporate financial analysis. A snowball approach was also applied to identify additional qualified experts.

The expert panel consisted of three groups: (1) accounting and finance university professors with relevant academic backgrounds; (2) capital market professionals from the Securities and Exchange Organization, Tehran Stock Exchange, investment companies, and brokerage firms; and (3) professional auditors experienced in corporate financial statement analysis. A total of 60 experts participated in the Fuzzy Delphi process, and the variables reaching the required consensus level were retained for subsequent modeling stages.

3.4 Identification of Financially Distressed Firms

Research data were collected from reports and disclosures published on CODAL. Unlike previous Iranian studies, which have considered financial distress and bankruptcy as equivalent concepts based on Article 141 of the Iranian Commercial Law, this study adopts a broader definition of financial distress. A firm was classified as financially distressed if it satisfied at least one of the following criteria:

- Negative net income or operating income before taxes for at least three consecutive years (Gilbert et al., 1990);
- Poor performance and retained losses for at least two consecutive years (Gilbert et al., 1990);
- Book value of equity lower than nominal capital for three consecutive years (Gilbert et al., 1990);
- Earnings before interest, taxes, depreciation, and amortization (EBITDA) lower than financial costs, or an EBITDA-to-financial-cost ratio of less than 80% for two consecutive years (Pindado, 2007);
- A decline of at least 40% in dividends per share (DPS) for three consecutive years (Denis & Denis, 1995);
- Negative operating cash flows for two or three consecutive years (Platt, 2010);
- Negative stock returns and sales growth for two consecutive fiscal years (Opler & Titman, 1994).

3.5 Text Mining Procedure

3.5.1 Document Conversion

The annual reports of the sample firms were converted from Acrobat PDF files into Microsoft Word documents using automated conversion tools. Companies whose reports did not meet the required quality after conversion were excluded from the sample.

3.5.2 Text Preprocessing

Given the importance of preprocessing in text mining, several procedures were carried out, including text normalization, tokenization, stop-word removal, word stemming, and the elimination of unnecessary words and characters.

Normalization: In this stage, all texts were standardized and normalized through script corrections.

Tokenization: The textual data were segmented into individual words and tokens to facilitate subsequent analyses.

Stop-word Removal: Stop words are terms with limited semantic value that occur frequently in documents, such as conjunctions and prepositions. Due to their high frequency and low informational content, these words were removed during preprocessing.

Word Stemming: Words sharing the same root were considered identical terms. For example, the terms "sale" and "sales" were treated as the same word.

Removal of Unnecessary Characters: All unnecessary symbols and characters, such as *, {}, @, and +, were eliminated from the textual data.

3.5.3 Feature Extraction and Selection

Following preprocessing, the Term Frequency–Inverse Document Frequency (TF-IDF) matrix was employed to determine the importance of each term and to select textual features. The TF-IDF score was calculated as follows:

$$TF-IDF_{w_{i,j}} = tf_{i,j} \times \log\left(\frac{N}{df}\right)$$

Where $tf_{i,j}$ represents the frequency of word w_i in text d_j , and df denotes the number of texts that include w_i . Moreover, N refers to the number of texts under analysis, and $w_{i,j}$ indicates the weight of w_i in text j .

In this method, the frequency of a word in a document indicates the importance of that term in documents, and the frequency of a document in a term (*i.e.*, df : the percentage of documents containing that term) indicates its importance across all texts. A low value of df means the uniqueness of this term in documents, as it does not appear in many documents. Therefore, idf was used instead of df for weighting in this study. As a result, a high weight in the $tf*idf$ method will indicate the high frequency of a word in a document and the few documents containing that text. In this step, the words with the highest score will determine the features.

3.6 Training the Model and Result Assessment and Comparison

At this stage, a combination of the word matrix, selected financial data, and the financial distress status of firms was used as the input of the models, followed by the implementation of machine learning algorithms to develop the prediction models. Specifically, Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Decision tree and k nearest neighborhood were employed to predict financial distress and compare their predictive performance.

The present study employs the following relationship to identify financial distress:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1n} \\ X_{21} & X_{22} & \dots & X_{2n} \\ X_{31} & X_{32} & \dots & X_{3n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{m1} & X_{m2} & \dots & X_{mn} \end{bmatrix} \times \begin{bmatrix} B_1 \\ B_2 \\ B_3 \\ \vdots \\ B_n \end{bmatrix}$$

The dependent variable (Y) is an $m \times 1$ matrix whose values are one for firms with financial distress and zero for non-distressed firms. The matrix of explanatory variables (X) is an $m \times n$ matrix, where m represents approximately 1,100 firm-year observations and n denotes the number of financial indicators and textual terms used to explain the financial distress index, among which those with the highest explanatory power are selected. Each matrix member demonstrates the relative frequency (TF-IDF) of a term. After developing this matrix, the study proceeded to the statistical modelling phase using the selected machine learning algorithms.

Various methods are available to assess the performance of learning models and classification results. Among these methods, the confusion matrix is one of the most widely used techniques for evaluating machine learning classification problems. The confusion matrix presents the classification results based on actual observations and displays the number of correct and incorrect predictions made in each class.

Based on the confusion matrix, the indices of Accuracy, Precision, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC) were calculated to evaluate and compare the predictive performance of the machine learning algorithms

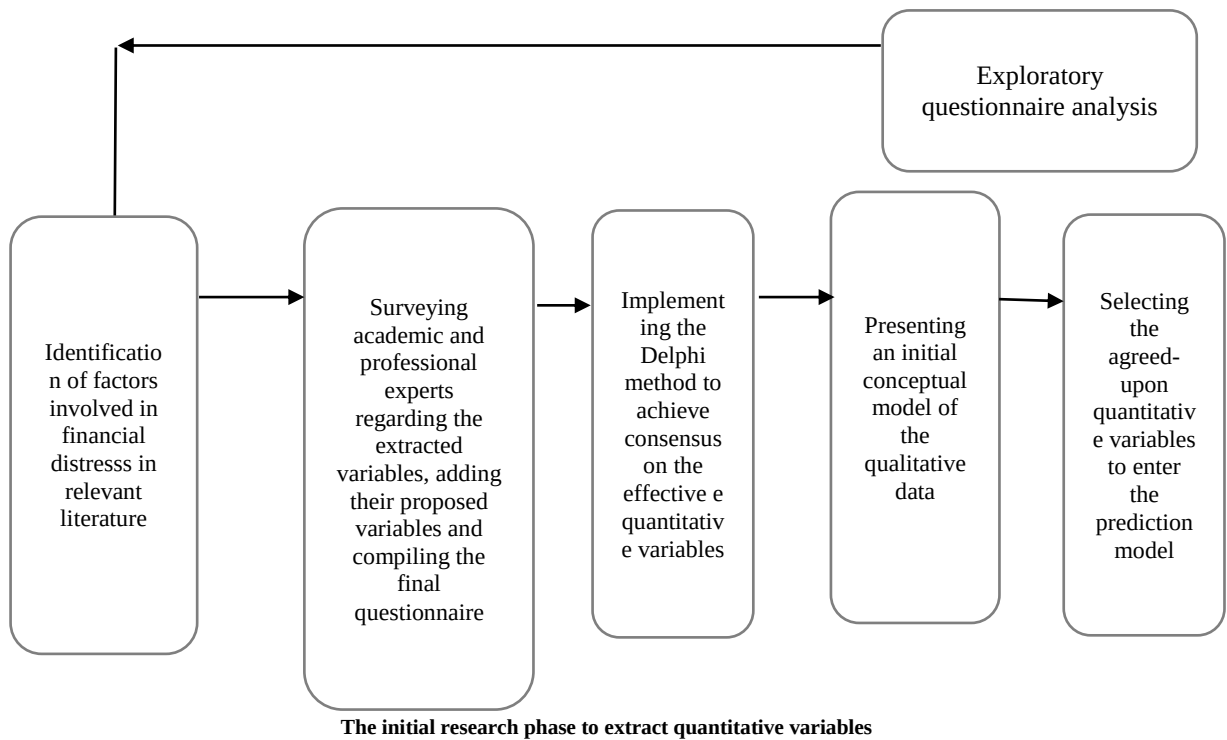


Figure 1
Extracting Quantitative Variables

The second research phase to extract qualitative variables

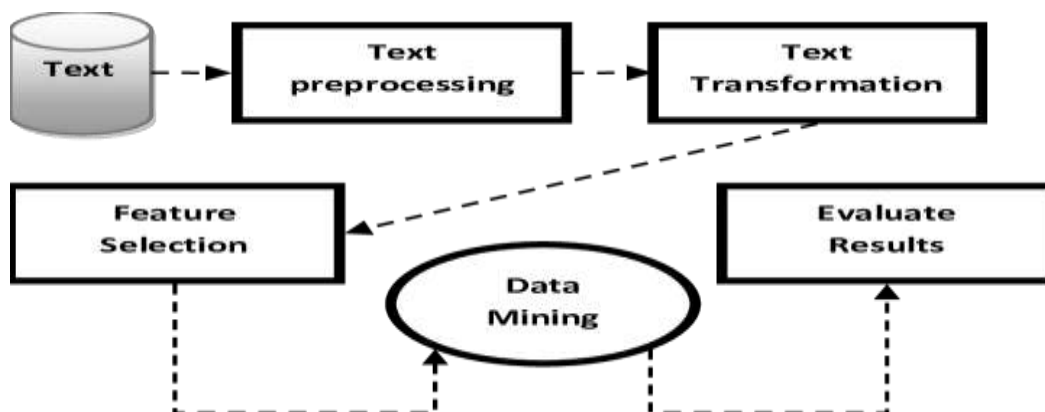


Figure 2
Extracting Qualitative Variables

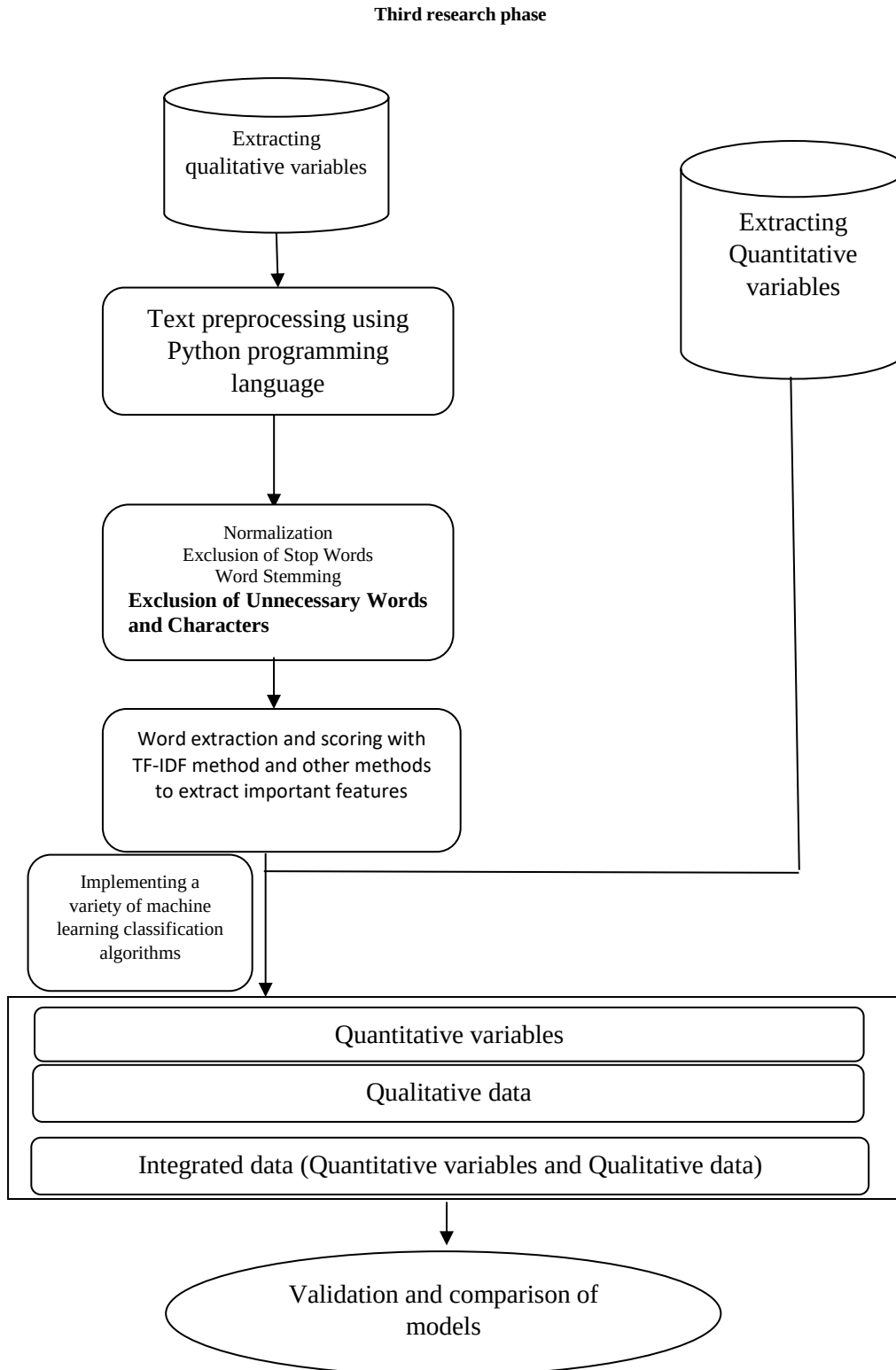


Figure 3
General Research Framework

4. Research findings

4.1. Descriptive statistics

To provide a better understanding of the characteristics of the sample firms, [Table 1](#) presents the descriptive statistics for a representative subset of the study variables, given the large number of quantitative variables included in the analysis.

Table1

Descriptive Statistics of Quantitative Variables

Variable	Mean	Median	Std. Dev.
Current Ratio	1.792	1.383	1.818
Net Income to Total Assets	0.141	0.112	0.156
Total Debt to Total Assets	0.558	0.559	0.223
Sales to Total Assets	0.931	0.802	0.561
Working Capital to Total Assets	0.186	0.189	0.241
Return on Equity (ROE)	0.186	0.283	2.416
Total Asset Growth	0.327	0.179	0.769
Current Assets to Total Assets	0.677	0.698	0.178

4.2 Empirical Results and Model Performance

The study of the research literature yielded a total of 62 financial distress variables that were provided to financial experts following the compilation of the questionnaire. Results of analyzing expert opinions using fuzzy Delphi indicated the significance of 53 quantitative indicators in financial distress as demonstrated in [Table 2](#).

Table 2

Results of Fuzzy Delphi to Extract Quantitative Variables

Row	Ratio	Status	De-fuzzified value
1	Current Ratio	Confirm	0.874
2	Return On Assets Ratio	Confirm	0.838
3	Debt Ratio	Confirm	0.889
4	Asset Turnover Ratio	Confirm	0.784
5	Working Capital to Total Assets	Confirm	0.77
6	Acid-test Ratio	Confirm	0.827
7	Return on Equity	Confirm	0.825
8	Receivables Turnover Ratio	Confirm	0.798
9	Gross Margin	Confirm	0.875
10	Operating Margin	Confirm	0.817
11	Net Margin	Confirm	0.878
12	Current Assets to Total Assets	Confirm	0.741
13	Interest Coverage	Confirm	0.804
14	Shareholders' Equity to Total Assets	Confirm	0.785
15	Growth of Total Assets	Confirm	0.704
16	Inventory Turnover Ratio	Confirm	0.744
17	Cash to Total Assets	Confirm	0.758
18	Interest Expense to Sales	Confirm	0.717
19	Cash to Current Liabilities	Confirm	0.765

Row	Ratio	Status	De-fuzzified value
20	Sale to Current Assets	Confirm	0.726
21	Current Debt to Total Assets	Confirm	0.779
22	Earnings Per Share	Confirm	0.721
23	Size (Logarithm of Book Value of Total Assets)	Reject	0.682
24	Sales to Working Capital	Confirm	0.748
25	Quick Assets to Total Assets	Confirm	0.74
26	Total Current Debt	Confirm	0.812
27	Cash to Net Sale	Confirm	0.74
28	Fixed Asset Returns	Confirm	0.727
29	Working Capital to Equity	Confirm	0.716
30	Net Income to Total Debt	Confirm	0.758
31	Current Liabilities to Equity	Confirm	0.775
32	Retained Earnings on Capital	Confirm	0.736
33	Loan to Capital	Confirm	0.777
34	EBIT/ Sale	Confirm	0.781
35	Operating Cash Flow to Total Assets	Confirm	0.821
36	Operating Cash Flow to Total Debt	Confirm	0.85
37	Operating Cash Flow to Current Liabilities	Confirm	0.835
38	Operating Cash Flow to Sales	Confirm	0.844
39	Operating Cash Flow to Equity	Confirm	0.749
40	Operating Cash Flow at Interest Expense	Confirm	0.844
41	Debt Coverage	Confirm	0.784
42	Dividend Per Share	Reject	0.701
43	Operating Cash Flow/ Cash Outflow for Investment and Ffinancing Activities	Confirm	0.733
44	Operating Cash Flow to Net Income	Confirm	0.802
45	Operating Cash Flow/Cash Paid for Long-Term Sssets	Confirm	0.709
46	Market Value of Shares to Book Value of Debt	Confirm	0.709
47	Market Value of Shares to Book Value of Shares	Reject	0.685
48	Size = Logarithm of Market Value	Reject	0.666
49	Stock Returns	Confirm	0.723
50	Stock Price	Reject	0.6
51	Risk	Confirm	0.778
52	Stock Trading Volume	Reject	0.648
53	Price to Earning Per Share	Confirm	0.704
54	GDP Growth	Confirm	0.717
55	Money Supply	Confirm	0.737
56	The Unemployment Rate	Reject	0.669
57	Total Stock Market Index	Reject	0.568
58	Dollar Market	Confirm	0.831
59	Inflation	Confirm	0.864
60	Oil Price Changes	Reject	0.636
61	Percentage Changes in the Average Interest Rate	Confirm	0.789
62	Operating Cash Flow to Operating Income	Confirm	0.729

Following the collection of quantitative research variables, the qualitative variables involved in financial distress were also identified through text mining.

As mentioned earlier, the preprocessing steps were aimed at preparing the texts of board activity reports for text mining operations. Therefore, Python was employed to perform the exclusion of stop words, word stemming, normalization, and exclusion of unnecessary words and characters. Figures 4, 5, 6, and 7 present some pseudocodes used in the foregoing steps.

```
import codecs
nmz = Normalizer()
stops = sorted(list(set([nmz.normalize(w) for w in codecs.open('non-verbal_stopwords.txt', encoding='utf-8').read().split('\n')
])))
stops1 = sorted(list(set([nmz.normalize(w) for w in codecs.open('persian5T.txt', encoding='utf-8').read().split('\n') if w
])))
stops2 = sorted(list(set([nmz.normalize(w) for w in codecs.open('stop-words.txt', encoding='utf-8').read().split('\n') if w
])))
stops3 = sorted(list(set([nmz.normalize(w) for w in codecs.open('stopwords_all.txt', encoding='utf-8').read().split('\n') if w
])))
stops4 = sorted(list(set([nmz.normalize(w) for w in codecs.open('verbal_stopwords.txt', encoding='utf-8').read().split('\n') if
])))
stops5 = sorted(list(set([nmz.normalize(w) for w in codecs.open('Persian_Stoplist.txt', encoding='utf-8').read().split('\n') if
])))
stops6 = sorted(list(set([nmz.normalize(w) for w in codecs.open('Persian_Stop_Words_List.txt', encoding='utf-8').read().split('\n')
])))
Stopword=stops+stops1+stops2+stops3+stops4+stops5+stops6
```

Figure 4

The Function to Remove Stop Words

```
def split_on_word(text):
    """Use regular expression tokenizer.
    Keep apostrophes.
    Returns a list of lists, one list for each sentence:
    [[word, word], [word, word, ..., word], ...].
    """
    if type(text) is list:
        return [regexp_tokenize(sentence, pattern="\w+(?:[-']\w+)*") for sentence in text]
    else:
        return regexp_tokenize(text, pattern="\w+(?:[-']\w+)*")
```

Figure 5

Converting Text Into Sentences And Then Into Words (Tokenization) And Stemming

```
def normalize(tokenized_words):
    """Removes stop words, numbers, short words, and lowercases text.
    Returns a list of lists, one list for each sentence:
    [[word, word], [word, word, ..., word], ...].
    """
    stop_words = Stopword
    return [[w.lower() for w in sent
              if (w.lower() not in stop_words)]
            for sent in tokenized_words]
```

Figure 6

Normalization Of Text (Normalize)

Text mining was performed after pre-processing the text, which yielded 5,911 terms whose scores were determined using the TF-IDF method –a technique to extract features from text. In general, term frequency (TF) in each document reveals the importance of a given term. On the other hand, the frequency of documents containing a given term (DF) demonstrates the significance of the given term across all documents. In other words, small DF indicates the exclusive status and rare usage of the term and its uniqueness across the documents. Therefore, this method considers inverse document frequency (IDF) to score terms. In this technique, every index in the matrix, which is the intersection of a row with a column, equals the result of dividing TF by IDF, representing the percentage of the repetitions of each specific term (columns) divided by the total term of a specific report (row).

Following the extraction of the dependent variable (Y matrix) besides the explanatory variables (X matrix) containing qualitative and quantitative variables, we entered the modeling phase using various machine learning methods (including logistic regression, random forest, decision tree, nearest neighborhood and Support vector machine (SVM) methods with linear, RBF, sigmoid and polynomial kernels) in Python and calculating the model's predictive power.

```
def clean_text(df, col):
    """A function for keeping only alpha-numeric
    characters and replacing all white space with
    a single space.
    """
    return df[col].apply(lambda x: re.sub('[\t\r\n]', ' ', x))\
        .apply(lambda x: re.sub(' ', ' ', x))\
        .apply(lambda x: re.sub('\n', ' ', x))\
        .apply(lambda x: re.sub('\t', ' ', x))\
        .apply(lambda x: re.sub('@', ' ', x))\
        .apply(lambda x: re.sub('/', ' ', x))\
        .apply(lambda x: re.sub('!', ' ', x))\
        .apply(lambda x: re.sub('?', ' ', x))\
        .apply(lambda x: re.sub('!', ' ', x))\
        .apply(lambda x: re.sub('\|', ' ', x))\
        .apply(lambda x: re.sub('\(', ' ', x))\
        .apply(lambda x: re.sub('\)', ' ', x))\
        .apply(lambda x: re.sub(':', ' ', x))\
        .apply(lambda x: re.sub(';', ' ', x)).strip())
```

Figure 7

Function Of Cleaning The Text From Punctuation Marks

Before delving into the process, an assessment of the observation frequency in each response group indicating whether or not financial distress occurred revealed 415 cases of financial distress and 658 cases of no financial distress. Given the significant gap between the two response variable groups, they were homogenized using the SMOTE sampling method, which left 658 samples in each group. Table 3 demonstrates the results of fitting the three methods.

Table 3

Results of Implementing Machine Learning Algorithms

Methode	Model	Specificity	Precision	Recall	F1 Score	AUC
Word	Random Forest	0.700	0.689	0.709	0.698	0.706
	Logistic Regression	0.776	0.764	0.800	0.780	0.787
	Decision Tree	0.799	0.804	0.778	0.790	0.788
	KNearest Neighborhood	0.683	0.630	0.761	0.688	0.722
	SVM with Linear Kernel	0.764	0.753	0.786	0.768	0.775
	SVM with Radial Kernel	0.826	0.816	0.863	0.838	0.845
	SVM with Polynomial Kernel	0.725	0.705	0.756	0.728	0.740
	SVM with Sigmoid Kernel	0.506	0.487	0.502	0.494	0.504
Fin	Random Forest	0.864	0.775	0.807	0.784	0.835
	Logistic Regression	0.822	0.698	0.757	0.714	0.789
	Decision Tree	0.837	0.727	0.791	0.750	0.814
	KNearest Neighborhood	0.864	0.762	0.882	0.808	0.873
	SVM with Linear Kernel	0.824	0.690	0.814	0.737	0.819
	SVM with Radial Kernel	0.936	0.893	0.888	0.885	0.911
	SVM with Polynomial Kernel	0.849	0.744	0.784	0.750	0.816
	SVM with Sigmoid Kernel	0.779	0.660	0.632	0.642	0.706
Fin+Word	Random Forest	0.858	0.744	0.930	0.821	0.893
	Logistic Regression	0.819	0.673	0.824	0.733	0.822
	Decision Tree	0.845	0.743	0.799	0.765	0.821
	KNearest Neighborhood	0.868	0.768	0.899	0.823	0.884
	SVM with Linear Kernel	0.781	0.660	0.636	0.644	0.707
	SVM with Radial Kernel	0.943	0.905	0.904	0.900	0.924
	SVM with Polynomial Kernel	0.864	0.761	0.876	0.808	0.870
	SVM with Sigmoid Kernel	0.781	0.660	0.636	0.644	0.709

K-fold cross validation was used in fitting each of the models. Model training and validation were conducted considering 10 folds for each sampling. As mentioned, accuracy recall (the percentage of correct identifications), F-1 score (a combination of accuracy and precision), and AUC as the area under the ROC curve were used to assess the predictive power of each machine learning model, the results of which were obtained from the average results of each fold.

As per the modeling results, the integrated method outperformed quantitative and qualitative methods. In other words, the mean F1 score of the integrated method was 77, whereas the mean F1 scores of quantitative and qualitative methods were 75 and 72, respectively. Moreover, the mean AUCs of integrated, quantitative, and qualitative methods were 83%, 82%, and 73%, respectively. The results also indicated the higher efficiency of the radial kernel SVM in detecting financial distress than the other models. Based on the AUC criterion, the detection power of the qualitative method was 85%, a figure to which the decision tree method came close with 79%. The detection power of the quantitative method was 91%, a figure to which the KNN method came close with 87%. Finally, the detection power of the integrated method was 92%, a figure to which the random forest method came close with 89%.

5. Results and Discussion

So far, many models have been presented for predicting financial distress, but what has been less discussed is the effect of the information content of the annual report on predicting financial distress. The report of the board of directors contains valuable clues that, like financial data, can detect the existence of financial distress in companies. Therefore, in this research, an attempt has been made to predict the financial distress of companies by using the information capacity of the annual reports. After preprocessing and data mining operations and extraction of dependent and independent variables, using different machine learning methods (including logistic regression, random forest, decision tree, nearest neighbor, and methods of vector machine with linear, RBF, sigmoid, and polynomial) methods, financial distress was measured. The prediction results indicate the high predictive power of SVM model with radial kernel compared to other models. So that its ability to predict financial distress based on three quantitative, qualitative and integrated methods is 85, 91 and 92% respectively. The findings of this study are consistent with those of Mai et al. (2019), Matin et al. (2019), and Tang et al. (2020), who reported that incorporating textual information significantly enhances the predictive power of financial distress models. The results of this research showed that instead of just paying attention to the numbers and the ratios derived from these numbers, the text mining technique can also be used for analysis and prediction, and by combining it with the results obtained from quantitative information, the financial distress of companies can be determined. Although the accuracy of predicting results from unstructured data compared to structured data (such as Wang et al., 2017, etc.) with a predictability about 90% is less, but knowing and using the capacity of this type of information is very important so many academics and practitioners believe that quantitative disclosures alone are not effective for economic decision-making (Guo, 2016).

In several ways, this research has increased knowledge of the field of accounting and data mining: First, the present research has tried to distance itself from the structured data extracted from financial statements, market data, and economic data, and for a major part of the analysis relies on the annual reports (unstructured data). Second, the results of this research instead of relying on "numbers and figures", relied on the analysis of "the text of reports" and predicted the financial distress of companies very accurately.

The current research has relied on the annual report to analyze the text. Meanwhile, other factors may also be suitable for predicting financial distress. Therefore, it is suggested that predictability of financial distress by using other similar qualitative reports, information such as the MD&A report, the text of the minutes of the board of directors' meetings, the text of the speeches of the board of directors and company managers, the text of telephone conferences, the text of the auditor's report, the text of the managers' letter, the text income announcements, and... to measure and evaluate. Researchers are also suggested to use other feature selection methods such as LDA, Doc2Vec to predict the company's financial distress and compare it with the results of this research.

The limitation of this research is related to the lack of accurate software to convert annual reports from PDF to WORD in Persian texts. Most of the available software for converting PDF to WORD format works well in English language, but in Persian language, due to the existence of various writing forms, it is currently impossible to cover all the grammatical states of words by applying a certain set of rules. Therefore, according to the progress of data science and the field of text mining in Persian language, the use of better and accurate software can increase the accuracy of input data and ultimately outputs.

References

- Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4), 589–609. <https://doi.org/10.1111/j.1540-6261.1968.tb00843.x>
- Altman, E. I., Iwanicz-Drozowska, M., Laitinen, E. K., & Suvas, A. (2017). Financial distress prediction in an international context: A review and empirical analysis of Altman's Z-score model. *Journal of International Financial Management & Accounting*, 28(2), 131–171. <https://doi.org/10.1111/jifm.12053>
- Barzegari Khanagha, J., Ansari Samani, H., & Razzazadeh, L. (2019). Predicting the financial status of companies using content analysis the reports of the board of directors. *Empirical Studies in Financial Accounting*, 16(64), 135–160. <https://doi.org/10.22054/qjma.2020.27663.1718>
- Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, 4, 71–111. <https://doi.org/10.2307/2490171>
- Denis, D. J., & Denis, D. K. (1995). Causes of financial distress following leveraged recapitalizations. *Journal of Financial Economics*, 37(2), 129–157. [https://doi.org/10.1016/0304-405X\(94\)00792-Y](https://doi.org/10.1016/0304-405X(94)00792-Y)

- FitzPatrick, P. J. (1932). A comparison of the ratios of successful industrial enterprises with those of failed companies. *The Certified Public Accountant*, 12, 598–605, 656–662, 727–731. <https://www.scrip.org/reference/referencespapers?referenceid=3100314>
- Gilbert, L. R., Menon, K., & Schwartz, K. B. (1990). Predicting bankruptcy for firms in financial distress. *Journal of Business Finance & Accounting*, 17(1), 161–171. <https://doi.org/10.1111/j.1468-5957.1990.tb00555.x>
- Guo, L., Shi, F., & Tu, J. (2016). Textual analysis and machine leaning: Crack unstructured data in finance and accounting. *The Journal of Finance and Data Science*, 2(3), 153–170. <https://doi.org/10.1016/j.jfds.2017.02.001>
- Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence*, 1(1), 60–76. <https://doi.org/10.4304/jetwi.1.1.60-76>
- Hájek, P., & Michalak, K. (2013). Feature selection in corporate credit rating prediction. *Knowledge-Based Systems*, 51, 72–84. <https://doi.org/10.1016/j.knosys.2013.07.008>
- Hajiha, Z., & Saberi Rochi, M. R. (2021). Text-mining technique and prediction of financial distress. *Accounting and Auditing Studies*, 10(38), 39–52. <https://doi.org/10.22034/iaas.2021.134533>
- Huang, Y., Wang, Z., & Jiang, C. (2024). Diagnosis with incomplete multi-view data: A variational deep financial distress prediction method. *Technological Forecasting and Social Change*, 201, 123269. <https://doi.org/10.1016/j.techfore.2024.123269>
- Jo, N. O., & Shin, K. S. (2016). Bankruptcy prediction modeling using qualitative information based on big data analytics. *Journal of Intelligence and Information Systems*, 22(2), 33–56. <https://doi.org/10.13088/jiis.2016.22.2.033>
- Karimpour, A., Azhdar, M., & Davoodi, S. M. R. (2025). Presenting a model for predicting financial distress based on deep learning. *Journal of Investment Knowledge*, 15(58), 417–439. <https://doi.org/10.30495/jik.2024.77615.4541>
- Ključnikov, A., Civelek, M., Krajčík, V., Novák, P., & Červinka, M. (2022). Financial performance and bankruptcy concerns of SMEs in their export decision. *Oeconomia Copernicana*, 13(3), 867–890. <https://doi.org/10.24136/oc.2022.025>
- Kovacova, M., Klietnik, T., Kubala, P., Valaskova, K., Radišić, M., & Borocki, J. (2018). Bankruptcy models: Verifying their validity as a predictor of corporate failure. *Polish Journal of Management Studies*, 18(1), 167–179. <https://doi.org/10.17512/pjms.2018.18.1.13>
- Li, F. (2010). Textual analysis of corporate disclosures: A survey of the literature. *Journal of Accounting Literature*, 29(1), 143–165. <https://ssrn.com/abstract=1756926>
- Li, S., Shi, W., Wang, J., & Zhou, H. (2021). A deep learning-based approach to constructing a domain sentiment lexicon: A case study in financial distress prediction. *Information Processing & Management*, 58(5), 102673. <https://doi.org/10.1016/j.ipm.2021.102673>
- Liu, J., & Jia, M. (2025). Financial distress prediction with annual reports-based deep textual feature extraction: A hybrid approach. *Information Sciences*, 686, 121318. <https://doi.org/10.1016/j.ins.2024.121318>
- Loughran, T., & McDonald, B. (2014). Measuring readability in financial disclosures. *The Journal of Finance*, 69(4), 1643–1671. <https://doi.org/10.1111/jofi.12162>
- Mai, F., Tian, S., Lee, C., & Ma, L. (2019). Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research*, 274(2), 743–758. <https://doi.org/10.1016/j.ejor.2018.10.024>
- Matin, R., Hansen, C., Hansen, C., & Mølgaard, P. (2019). Predicting distresses using deep learning of text segments in annual reports. *Expert Systems with Applications*, 132, 199–208. <https://doi.org/10.1016/j.eswa.2019.04.071>
- Mehrani, S., Mehrani, K., & Karami, G. R. (2005). The use of historical financial and non-financial information to separate winner firms from loser firms. *The Iranian Accounting and Auditing Review*, 11(38), 77–92. <https://doi.org/10.1001.1.26458020.1383.11.4.3.9>
- Merwin, C. L. (1942). *Financing small corporations in five manufacturing industries, 1926–36*. National Bureau of Economic Research. <https://ideas.repec.org/b/nbr/nberbk/merw42-1.html>
- Nguyen, B. H., Nguyen, H. T. T., & Huynh, V. N. (2026). Enhancing corporate bankruptcy prediction with machine learning and textual analysis. In *Data science in finance and accounting* (pp. 133–159). Springer. https://doi.org/10.1007/978-3-032-06179-9_7
- Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18(1), 109–131. <https://www.jstor.org/stable/2490395>
- Opler, T. C., & Titman, S. (1994). Financial distress and corporate performance. *The Journal of Finance*, 49(3), 1015–1040. <https://doi.org/10.1111/j.1540-6261.1994.tb00086.x>
- Pindado, J., Rodrigues, L., & De la Torre, C. (2008). Estimating financial distress likelihood. *Journal of Business Research*, 61(9), 995–1003. <https://doi.org/10.1016/j.jbusres.2007.10.006>
- Platt, H. D., & Platt, M. B. (2006). Understanding differences between financial distress and bankruptcy. *Review of Applied Economics*, 2(2), 141–157. <https://doi.org/10.22004/ag.econ.50146>
- Sreedharan, M., Khedr, A. M., & El Bannany, M. (2020). A comparative analysis of machine learning classifiers and ensemble techniques in financial distress prediction. In *2020 17th International Multi-Conference on Systems, Signals & Devices (SSD)* (pp. 653–657). IEEE. <https://doi.org/10.1109/SSD49366.2020.9364178>

- Szymura, A. (2022). Risk assessment of Polish joint stock companies: Prediction of penalties or compensation payments. *Risks*, 10(5), 102. <https://doi.org/10.3390/risks10050102>
- Tang, X., Li, S., Tan, M., & Shi, W. (2020). Incorporating textual and management factors into financial distress prediction: A comparative study of machine learning methods. *Journal of Forecasting*, 39(5), 769–787. <https://doi.org/10.1002/for.2661>
- Wang, C., Gong, P., Li, J., & Wang, Z. (2025). Corporate financial distress prediction with multiperiod annual report data: A fusion deep neural network model. *PLOS ONE*, 20(9), e0333064. <https://doi.org/10.1371/journal.pone.0333064>
- Wang, Z., Chen, M. H., Chin, C. L., & Zheng, Q. (2017). Managerial ability, political connections, and fraudulent financial reporting in China. *Journal of Accounting and Public Policy*, 36(2), 141–162. <https://doi.org/10.1016/j.jaccpubpol.2017.02.004>
- Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research*, 22, 59–82. <https://doi.org/10.2307/2490859>